


RESEARCH ARTICLE

Open Access



# Using random forest to predict antimicrobial minimum inhibitory concentrations of nontyphoidal *Salmonella* in Taiwan

Chia-Chi Wang<sup>1†</sup>, Yu-Ting Hung<sup>2,3†</sup>, Che-Yu Chou<sup>4</sup>, Shih-Ling Hsuan<sup>3</sup>, Zeng-Weng Chen<sup>2</sup>, Pei-Yu Chang<sup>5</sup>, Tong-Rong Jan<sup>1\*</sup> and Chun-Wei Tung<sup>4,5\*</sup> 

## Abstract

Antimicrobial resistance (AMR) is a global health issue and surveillance of AMR can be useful for understanding AMR trends and planning intervention strategies. *Salmonella*, widely distributed in food-producing animals, has been considered the first priority for inclusion in the AMR surveillance program by the World Health Organization (WHO). Recent advances in rapid and affordable whole-genome sequencing (WGS) techniques lead to the emergence of WGS as a one-stop test to predict the antimicrobial susceptibility. Since the variation of sequencing and minimum inhibitory concentration (MIC) measurement methods could result in different results, this study aimed to develop WGS-based random forest models for predicting MIC values of 24 drugs using data generated from the same laboratories in Taiwan. The WGS data have been transformed as a feature vector of 10-mers for machine learning. Based on rigorous validation and independent tests, a good performance was obtained with an average mean absolute error (MAE) less than 1 for both validation and independent test. Feature selection was then applied to identify top-ranked 10-mers that can further improve the prediction performance. For surveillance purposes, the genome sequence-based machine learning methods could be utilized to monitor the difference between predicted and experimental MIC, where a large difference might be worthy of investigation on the emerging genomic determinants.

**Keywords** Antimicrobial resistance, *Salmonella*, minimum inhibitory concentrations, machine learning, k-mer, whole-genome sequencing

<sup>†</sup>Chia-Chi Wang and Yu-Ting Hung contributed equally to this work.

Handling editor: Freddy Haesebrouck

\*Correspondence:

Tong-Rong Jan

tonyjan@ntu.edu.tw

Chun-Wei Tung

cwtung@nhri.edu.tw

<sup>1</sup> Department and Graduate Institute of Veterinary Medicine, School of Veterinary Medicine, National Taiwan University, Taipei 106, Taiwan

<sup>2</sup> Animal Technology Laboratories, Agricultural Technology Research Institute, Hsinchu City 300, Taiwan

<sup>3</sup> Graduate Institute of Veterinary Pathobiology, College of Veterinary Medicine, National Chung Hsing University, Taichung 402, Taiwan

<sup>4</sup> Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei 106, Taiwan

<sup>5</sup> Institute of Biotechnology and Pharmaceutical Research, National Health Research Institutes, Miaoli County 350, Taiwan



## Introduction

*Salmonella*, belonging to the Enterobacteriaceae family, is a gram-negative rods bacillus. They are widely distributed in animals and are prevalent in food-producing animals, including cattle, porcine, and poultry [1]. *Salmonella* is one of the major food-borne zoonotic pathogens causing approximately 93.8 million global infections with 155 000 deaths per year [2]. The symptoms of salmonellosis are generally mild; however, the severity of the disease depends on the serotypes of *Salmonella* and host factors [3, 4]. The World Health Organization (WHO) estimates that *Salmonella* is one of the four key causes of global diarrhoeal diseases. *Salmonella* has been considered the first priority for inclusion in a program of integrated surveillance of antimicrobial resistance (AMR) in foodborne pathogens by the WHO [5]. In Taiwan, *Salmonella* isolated from healthy poultry and swine is one of the major detected pathogens in our national AMR surveillance program from 2017.

AMR poses a major health problem worldwide. It was estimated that 4.95 million deaths were associated with bacterial AMR in 2019 [6]. Surveillance programs at multidisciplinary level are vital for better understanding of AMR and minimizing the emergence of AMR [7]. Recent advances in fast and affordable whole-genome sequencing technologies have revolutionized microbial surveillance [8]. Whole-genome sequence-based surveillance enabled the detection of multidrug-resistance (MDR) and can be a replacement for phenotypic tests [9–11]. While various online tools and databases are available for AMR detection and surveillance [8], the conventional methods are based on the search of AMR genes using a curated knowledge base such as the Comprehensive Antibiotic Resistance Database (CARD) [12] and ResFinder [13]. However, a knowledge gap was found that may impede the adoption of the conventional methods [10].

Machine learning algorithms are emerging tools for identifying AMR. The genomic sequences can be encoded as feature vectors for training prediction models of AMR. In contrast to the conventional methods relying on sequence comparison to a database, the machine learning methods learn patterns of AMR from training dataset without the issue of knowledge gap and usually provide superior performance over the conventional methods [14, 15]. The machine learning methods and software for AMR detection have been comprehensively reviewed [14]. The methods can be generally classified into qualitative and quantitative methods. The qualitative methods predict AMR based on a predefined dataset of susceptible and resistant isolates. Since the breakpoints for defining AMR may change, retraining of the qualitative methods will be required. Furthermore, the classification of isolates with minimum inhibitory concentration

(MIC) near the breakpoint could be unreliable and the isolates were often excluded from the development of qualitative methods [15]. In contrast, the quantitative methods directly predict the MIC values that provide a more flexible option for interpreting the prediction results [16, 17].

Since an MIC variation of up to two two-fold dilutions across laboratories was observed [15, 18, 19], the development and deployment of machine learning algorithms for MIC prediction can benefit from a dataset with well-controlled experimental conditions. In this study, random forest models were developed for AMR prediction of nontyphoidal *Salmonella*. Random forest is a popular ensemble tree-based algorithm consisting of multiple trees, each learned from different bagging samples. The average of predicted values from all trees will be the final predicted result. It is robust even for a small dataset and capable of dealing with high-dimensionality [20], that is suitable for the present work. The sequencing data and MIC measurement for 24 drugs were all generated from the antimicrobial surveillance program supported by the Bureau of Animal and Plant Health Inspection and Quarantine in Taiwan using the same protocols and conducted in the same laboratories. The robust prediction on new isolates not involved in developing the models showed the effectiveness of the models.

## Materials and methods

### Dataset

The WGS and MIC data were obtained based upon works supported by the Council of Agriculture, Executive Yuan, Taiwan, ROC, under grant numbers 106AS-9.12.1-BQ-B1, 107AS-8.9.1-BQ-B1, 108AS-8.8.1-BQ-B1, 109AS-8.8.1-BQ-B1, 110AS-5.6.1-BQ-B1, and 111AS-5.6.1-BQ-B1. The *Salmonella* strains were isolated from fecal samples collected randomly from healthy poultry and swine in slaughterhouses in Taiwan. A total of 321 *Salmonella* isolates collected before 2020 were utilized for model training and validation. For each drug, the associated isolates were divided into a training and a validation dataset in a ratio of 8:2. Additional 16 *Salmonella* isolates collected in 2020 were utilized as independent test dataset for assessing the prediction performance of the developed model. The WGS reads were generated using an Illumina MiSeq sequencer (Illumina®, San Diego, CA, USA) with paired-end 150 bp sequencing. The reads were trimmed at a Phred quality score of Q30 using Trimmomatic [21], and were de novo assembled using Unicycler with an Illumina-only assembly pipeline [22]. Parameters for genomes assembling include a minimum length of 75 bp. The genome assemblies were used for further analysis.

The final genome assemblies were checked by the quality metrics of the depth of coverage, total read length, N50 and number of contigs for its contiguity. Genome completeness was measured by alignment search of expected gene content and reference genome in *Salmonella* In Silico Typing Resource (SISTR) database. We followed the quality assessment recommended by the EU Reference Laboratory for antimicrobial resistance (EURL-AR) [23] and revised the quality standard based on our previous experience in strain identification, subtyping, and phylogenetic analysis. The good quality was set at over 50-fold depth of coverage, 4.5–5.5 Mb of total read length, over 20 000 bp of N50 and less than 1000 contigs. If the genome did not reach the quality standard as mentioned above, it was regarded as low quality. Low-quality genomes were removed from all further analysis and sequencing was carried out again.

The MICs of 24 antimicrobial agents, including amoxicillin, ampicillin, azithromycin, cefotaxime, cefoxitin, ceftazidime, ceftiofur, ceftriaxone, chloramphenicol, ciprofloxacin, colistin, enrofloxacin, ertapenem, florfenicol, gentamicin, kanamycin, meropenem, nalidixic acid, oxytetracycline, streptomycin, sulfonamide, tetracycline, tigecycline, and trimethoprim for *Salmonella* isolates were determined by broth microdilution method in accordance with the guideline of the Clinical and Laboratory Standards Institute (CLSI, USA) [24]. The antimicrobial agents were tested at two-fold dilution series with a maximum concentration from 64 to 1024 µg/mL. There are some isolate-drug pairs without MIC evaluation leading to a total number of 4924 and 1246 isolate-drug pairs for model training and validation, respectively. For the independent test dataset, only MICs for 11 drugs were evaluated for the 16 isolates resulting in 176 isolate-drug pairs (shown in Table 1). The breakpoints of CLSI (2021) for the 24 drugs were utilized for classifying

susceptible and resistant isolates. The  $\log_2$ -transformed MIC ( $\log_2$ MIC) values were utilized for following analysis. Detailed numbers of the datasets were shown in Table 2.

### Model development and feature selection

In this study, 10-mer features extracted from the genome sequences of isolates were utilized for machine learning. The k-mer counting (KMC) program [25] was applied to calculate the 10-mer frequencies as features. Theoretically, there will be  $4^{10}=1\ 048\ 576$  features. After removing the 10-mers not found in our dataset, the total number of features based on the 321 isolates of the training and validation datasets is 524 301. Please note that the test dataset might have additional 10-mer features not found in the training and validation dataset, but those 10-mers were ignored in this experiment. The command utilized for calculating 10-mer counts is “kmc -k 10 -fm -ci1 -cs1677215 input output temp”.

The random forest algorithm was applied to train a prediction model for  $\log_2$ MIC. The random forest algorithm has been shown to be effective for predicting AMR [26]. Random forest is an ensemble of  $n$  decision trees trained on bootstrap samples and  $m$  randomly selected features. The prediction results were the average of the predictions from the tree ensembles, where the decimals were rounded off. The number of randomly selected features for tree building was set to a default value of  $m=724$  that is the square root of the total feature number. The parameter of  $n$  was tuned based on out-of-bag (OOB) MAE. The OOB error is a method to estimate predictive performance by applying the model to predict OOB samples that were not involved in the development of a tree. In this study, we considered the  $n \in \{100, 200, 500, 700, 1000, 1500\}$ . The random seed was set to 0 for reproducibility.

**Table 1** The numbers of susceptible and resistant isolates based on the CLSI breakpoints (2021) in the independent test dataset for 11 drugs.

Drugs	Resistant	Susceptible	$\log_2$ MIC (Min., Q1, Q2, Q3, Max.)
Amoxicillin	15	1	(0, 6, 8, 8, 8)
Ceftiofur	0	16	(0, 0, 1, 2, 4)
Chloramphenicol	13	3	(3, 7.75, 8, 8, 8)
Ciprofloxacin	1	15	(-2, -1.25, -1, -1, 3)
Colistin	1	15	(-1, -3, -1.5, 1, 2)
Enrofloxacin	0	16	(-1, 0, 0, 0, 1)
Florfenicol	12	4	(4, 5.5, 8, 8, 8)
Gentamicin	1	15	(-1, -1, 0, 1, 6)
Kanamycin	9	7	(2, 4, 8, 8, 8)
Nalidixic acid	7	9	(3, 4, 4, 5.75, 8)
Oxytetracycline	14	2	(3, 8, 8, 8, 8)

**Table 2** The numbers of susceptible and resistant isolates in the training and validation dataset and clinical breakpoints for 24 drugs.

Drugs	Training	Resistant	Susceptible	log <sub>2</sub> MIC (Min., Q1, Q2, Q3, Max.)	Validation	Resistant	Susceptible	log <sub>2</sub> MIC (Min., Q1, Q2, Q3, Max.)	Clinical Breakpoint (µg/mL)
Amoxicillin	256	215	41	(-1, 8, 8, 8, 8)	65	54	11	(-1, 8, 8, 8, 8)	32
Ampicillin	167	135	32	(0, 8, 8, 8, 8)	42	33	9	(0, 8, 8, 8, 8)	32
Azithromycin	146	21	125	(1, 1, 1.5, 2, 8)	37	5	32	(1, 1, 2, 3, 8)	32
Cefotaxime	167	42	125	(-3, -3, -2, 1, 6)	42	10	32	(-3, -3, -2, 1.25, 6)	4
Cefoxitin	146	46	100	(1, 2, 3, 6, 8)	37	11	26	(-3, 2, 3, 7, 8)	32
Ceftazidime	167	27	140	(-1, -1, 0, 1, 8)	42	7	35	(-1, 0, 0, 1, 8)	16
Ceftiofur	256	23	233	(-1, 0, 0, 1, 8)	65	6	59	(-1, 0, 1, 1, 8)	32
Ceftriaxone	167	27	140	(-3, -3, -3, -2, 6)	42	6	36	(-3, -3, -2, -1.75, 4)	4
Chloramphenicol	256	178	78	(1, 3, 8, 8, 8)	65	45	20	(1, 4, 8, 8, 8)	32
Ciprofloxacin	256	27	229	(-7, -5, -3, -2, 4)	65	7	58	(-7, -3, -2, -1, 6)	1
Colistin	256	44	212	(-3, -3, -1, 1, 3)	65	12	53	(-3, -3, -1, 1, 6)	4
Enrofloxacin	256	2	254	(-1, -1, -1, -1, 5)	65	1	64	(-1, -1, -1, 0, 7)	32
Ertapenem	146	14	132	(-3, -3, -3, -3, 6)	37	3	34	(-3, -3, -3, -3, 6)	2
Florfenicol	256	161	95	(1, 3, 6, 8, 8)	65	41	24	(1, 3, 7, 8, 8)	32
Gentamicin	256	34	222	(-1, -1, -1, 1, 8)	65	9	56	(-1, -1, -1, 1, 8)	16
Kanamycin	256	73	183	(0, 1, 2, 8, 8)	65	19	46	(0, 1, 2, 8, 8)	64
Meropenem	167	1	166	(-3, -3, -3, -3, 6)	42	0	42	(-3, -3, -3, -3, -2)	4
Nalidixic Acid	256	102	154	(1, 2, 4, 8, 8)	65	26	39	(1, 2, 4, 8, 8)	32
Oxytetracycline	256	215	41	(0, 7, 8, 8, 8)	65	54	11	(0, 7, 8, 8, 8)	32
Streptomycin	167	122	45	(1, 4, 5, 8, 8)	42	30	12	(1, 4.75, 5, 8, 8)	32
Sulfonamide	167	119	48	(5, 7, 10, 10, 12)	42	29	13	(3, 9.75, 10, 10, 10)	512
Tetracycline	167	135	32	(-1, 6, 7, 8, 8)	42	34	8	(-1, 6, 7, 8, 8)	16
Tigecycline	167	0	167	(-3, -3, -2, -1, 0)	42	0	42	(-3, -2, -2, -1, -1)	4
Trimethoprim	167	115	52	(-1, -1, 8, 8, 8)	42	29	13	(-1, 6.25, 8, 8, 8)	16

For feature selection, the built-in feature importance function was utilized to rank the features for their importance. Subsequently, the top  $k$  features were adopted for model training and OOB error evaluation, where  $k \in \{100, 200, \dots, 2000\}$ . The scikit-learn 0.23.1 library and python 3.7 programming language were utilized to implement the random forest regressor.

#### Performance measurement

The present work aims to predict the MIC value for isolates using genome sequence, therefore the main indicator for measuring the performance of models is the mean absolute error (MAE). As for the classification results based on clinical breakpoints, accuracy, sensitivity,

specificity and precision were utilized for evaluating model performance as shown in the following:

$$\text{Accuracy} = \frac{TP + TN}{N} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

where  $N$ ,  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  represent the total number of samples, true positives, true negatives, false positives, and false negatives, respectively.

## Results

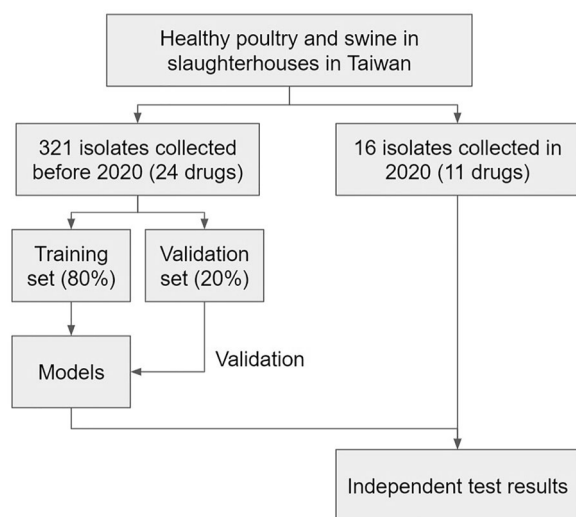
### Model development, validation and independent test

For model development, parameter tuning was conducted using the training dataset and the tuned parameter was then utilized for training the final prediction models. The system flow of this study is shown in Figure 1. A total of 24 models were developed for predicting the MICs of 24 drugs. The tree number of  $n=1000$  gave the best performance on OOB samples for 24 drugs with an average MAE of 0.916 (Figure 2A). The MAE ranging from 0.916 ( $n=1000$ ) to 0.927 ( $n=100$ ) for various numbers of trees indicates a small effect of the parameter of tree number on the MAE performance. The MAE value of less than 1 means the prediction will generally fall within a two-fold dilution range. When taking the breakpoint into consideration, the random forest models are able to distinguish susceptible and resistant isolates with an average accuracy of 91% (Figure 2B). Detailed OOB performance was shown in Additional file 1. The drugs meropenem, tigecycline, and enrofloxacin are associated with the lowest MAEs of less than 0.6. In contrast, ertapenem, kanamycin, and cefoxitin are associated with the worst MAEs of greater than 1.2. To have a better insight, the predictions were further evaluated by using measurements of sensitivity, specificity and precision. Please note that some drugs are associated with only a few resistant isolates that are expected to have low sensitivity. There are 11 drugs with a sensitivity higher than 0.8. Since there is no resistant isolate for tigecycline in the

dataset, sensitivity was not calculated. High specificity was obtained for all drugs except for streptomycin with a specificity of less than 0.69. High precision of greater than or equal to 0.8 was obtained for 18 drugs. For three drugs of enrofloxacin, meropenem and tigecycline, all isolates were predicted to have MIC values less than the breakpoints and therefore there are no calculated precision values.

To further evaluate the 24 models, the model performance on the validation dataset was shown in Figure 2 and Additional file 2. Please note that the validation dataset was not involved in the model training, therefore it represents a test on unseen isolates. Overall, similar average values of MAEs and accuracies of 0.92 and 0.92 were obtained for the validation dataset, respectively. A total of 11, 23 and 18 drugs are associated with a performance value greater than or equal to 0.8 in terms of sensitivity, specificity and precision, respectively. The performance measures based on OOB and validation dataset are very similar and are therefore considered less overfitting issues. Four drugs of meropenem, tigecycline, enrofloxacin, and sulfonamide are associated with the lowest MAEs of less than 0.6. Five drugs of ertapenem, cefoxitin, ciprofloxacin, tetracycline, and gentamicin are associated with worst MAEs of greater than 1.2.

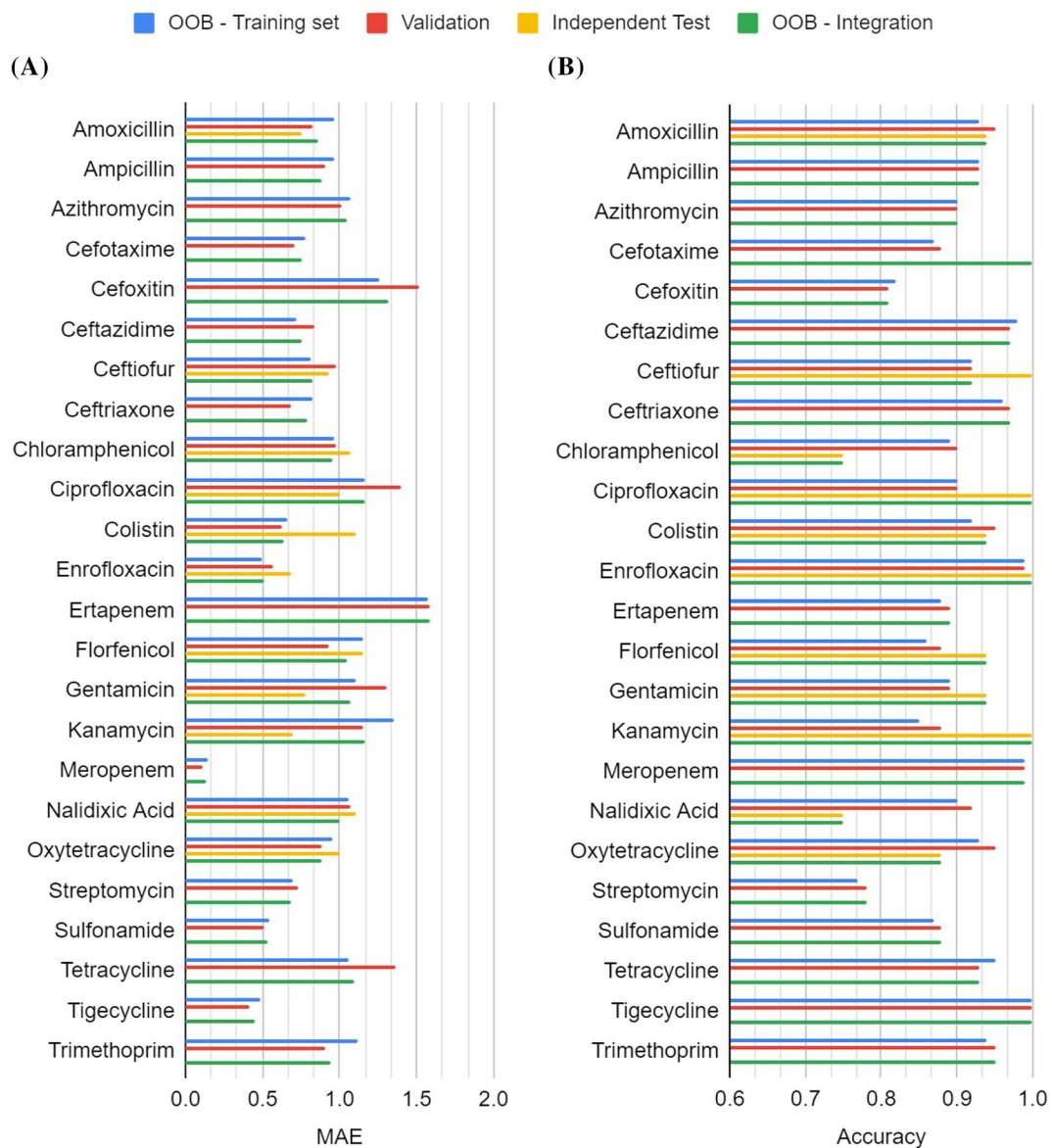
While the models provide good performance on OOB and validation datasets, an additional dataset collected in 2020 was utilized to independently test the models to simulate the application of the models for AMR surveillance. The MIC values of 176 unseen isolate-drug pairs were predicted based on the above-mentioned models. As shown in Figure 2, The average MAE and accuracy of 0.94 and 0.92, respectively, are similar to the results obtained from OOB and validation showing no overfitting problems. Detailed information is shown in Additional file 3.



**Figure 1** System flow of the present study.

### Performance improvement by enlarging dataset

While the developed prediction models gave a good prediction of MIC values for unseen isolates, the training dataset is relatively small. Since dataset size is a critical factor for developing machine learning models, the continuous integration of newly sequenced and phenotyped isolates into the training dataset could benefit the models. Therefore, this study evaluated the performance change made by enlarging the training dataset. An integrated training dataset was developed by integrating the original training, validation and test datasets. The integration resulted in a dataset of 6346 isolate-drug pairs that were utilized to train new models and evaluate the corresponding OOB errors. As expected, the integration of 1422 isolate-drug pairs



**Figure 2** Performance comparison for 24 drugs and four datasets of training, validation, test and integrated datasets. OOB out-of-bag, MAE mean absolute error.

improved the average OOB MAE and accuracy by 3.05% and 1.01% with values of 0.88 and 0.92, respectively. A comparison of the OOB performance using the training and integrated datasets is shown in Figure 2. Detailed performance was shown in Additional file 4. The largest improvement in MAE was made for kanamycin, trimethoprim, and amoxicillin with 0.19, 0.18 and 0.11 improvements, respectively. As for accuracy, 15%, 13% and 10% improvement was obtained for kanamycin, cefotaxime, and ciprofloxacin, respectively.

Future integration of more isolate-drug pairs could further improve the performance.

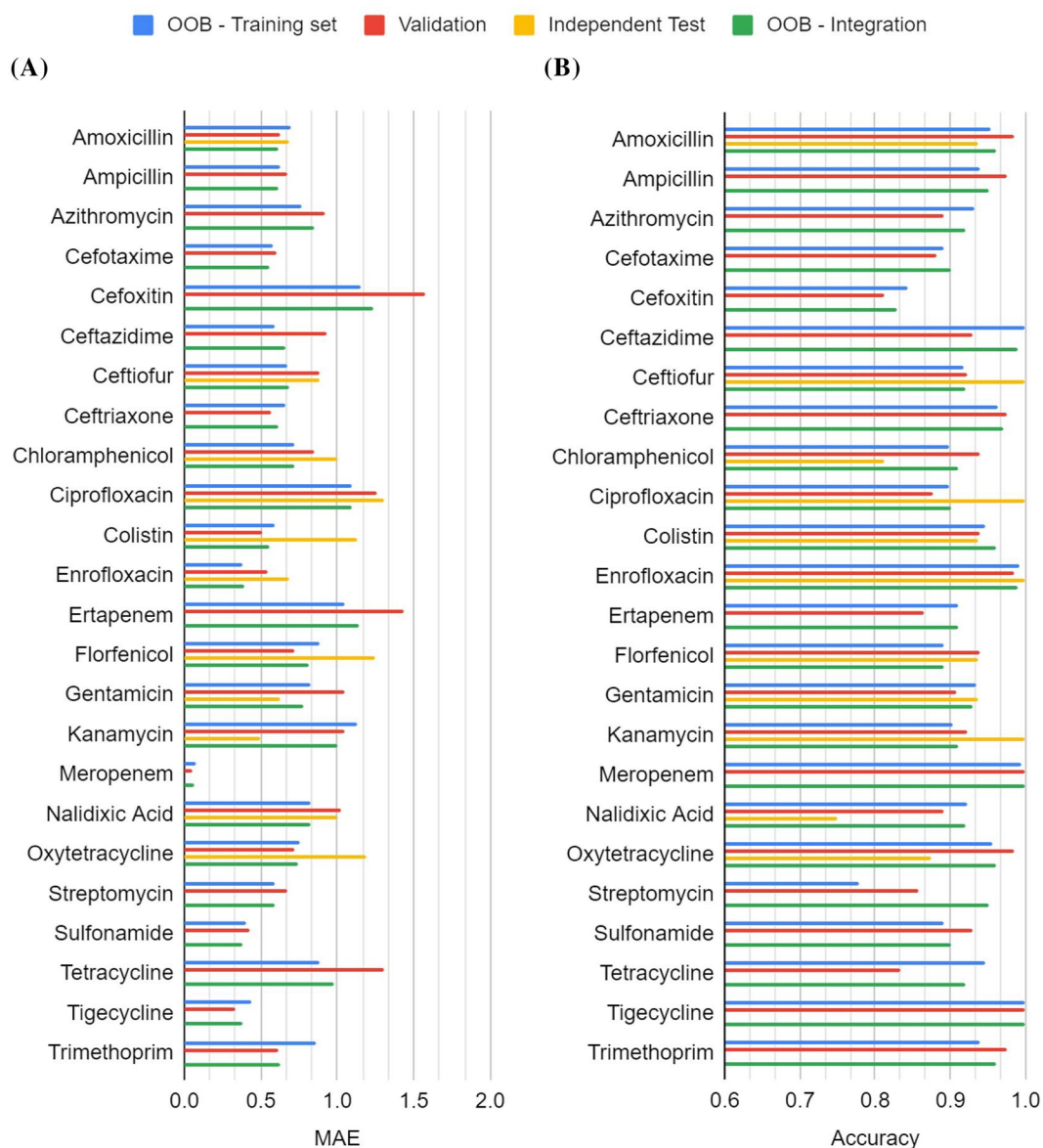
#### Top-ranked 10-mers as predictive features

The developed models are predictive, however, the high-dimensional feature vector could interfere with the performance of the applied machine learning algorithm and slow down the execution time. The top- $k$  features ranked by using the built-in feature importance function estimator of random forest with the lowest OOB and MAE

values were identified for each drug. Random forest algorithms were then applied to develop prediction models using the training dataset and top-*k* features. Selected top-ranked 10-mers for each drug are shown in Additional file 5. As shown in Additional file 6, the average MAE and accuracy of 0.72 and 0.93, respectively, were obtained that were much better than the models utilizing all 10-mer features. None of the drugs has an MAE value greater than 1.2. Four drugs of enrofloxacin, meropenem, sulfonamide and tigecycline are associated with low MAE values of less than 0.6. When applying the top-10 models to the validation dataset, the average MAE and accuracy

were 0.81 and 0.93, respectively (Additional file 7). Both results suggested that the top-ranked 10-mers are essential predictive features for MIC prediction.

An independent test on the additional dataset consisting of 176 unseen isolate-drug pairs collected in 2020 showed slightly improved performance (1%). The average MAE and accuracy of the model using the top-ranked 10-mers are 0.93 and 0.93, respectively. Detailed information is shown in Additional file 8. Compared to the large improvement on the validation dataset, it is unexpected that only a small improvement was made on the independent test dataset. As shown in Figure 3A, the

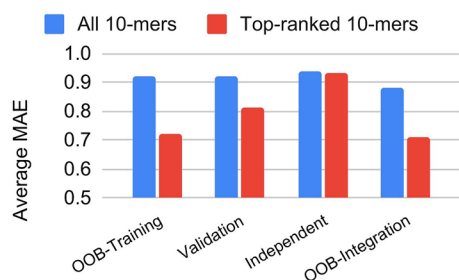


**Figure 3** Performance comparison for 24 drugs and four datasets of training, validation, test and integrated datasets using top-ranked 10-mers. *OOB* out-of-bag, *MAE* mean absolute error.

independent test performance of MAE for colistin, florfenicol and oxytetracycline are much worse than those observed in OOB and validation datasets. In contrast, chloramphenicol and nalidixic acid showed much worse accuracies. Since there are only 16 isolates per drug, the bias in performance may be introduced by the small dataset. In addition, gene mutation may result in different 10-mer profiles that may not be captured by the model. To incorporate all available information for developing final models, the integrated dataset of training validation and independent test datasets was utilized to train the final models using the selected top-ranked 10-mers. Detailed information was shown in Additional file 9. The average MAE and accuracy of the model using the top-ranked 10-mers are 0.71 and 0.94, respectively. Altogether, as shown in Figure 4, the top-ranked 10-mers and integration of additional datasets are useful for improving the prediction of MIC.

#### Comparison to existing methods

To provide a comparison of the developed method and existing methods, a publicly available machine learning-based tool [16] trained on a public database of PATRIC [27] and a knowledge-based method of ResFinder were applied to predict the samples of the independent test dataset. The machine learning-based tool utilizing XGBoost algorithm was reported to achieve high accuracy for predicting MICs for 15 antibiotics. However, the application of the models for predicting samples of the independent test dataset showed relatively low performance with an average MAE of 2.929 and accuracy of 65%, respectively. Note that some of the antibiotics in the independent test dataset were not covered by the XGBoost models and only amoxicillin, ceftiofur, chloramphenicol, ciprofloxacin, gentamicin, kanamycin, and nalidixic acid were considered in this comparison. For the ResFinder 4.1, the assembled contigs were submitted to the web server at with default parameters. The average accuracy for predicting the samples of the independent test dataset is 75% for amoxicillin, ceftiofur,



**Figure 4** Comparison of MIC prediction based on all k-mers and top-ranked k-mers.

chloramphenicol, ciprofloxacin, colistin, florfenicol, gentamicin, kanamycin, and nalidixic acid. Detailed performances for the XGBoost-based model and ResFinder are available as Additional files 10 and 11. Both methods did not provide satisfactory performance and stress out the importance of the present work.

#### Discussion

AMR is a global health issue and surveillance of AMR can be useful for understanding AMR trends and planning intervention strategies. However, traditional antimicrobial susceptibility testing is time-consuming and labor-intensive. Modern artificial intelligence methods are capable of capturing the patterns hidden in the dataset generated by previous experiments and applying the patterns for predicting unseen samples. As the inter-laboratory variation of MIC measurement can vary up to two two-fold dilutions, the publicly available tool [16] trained on a public database of PATRIC [27] did not produce good prediction results in our datasets. Also, the knowledge-based method of ResFinder did not provide satisfactory performance. Considering the possible knowledge gap and variation of sequencing and MIC measurement methods, it is desirable to develop genome sequence prediction models based on the data produced using the same protocols by the same laboratories in Taiwan.

In this study, genome sequence-based random forest models were developed for predicting MIC values of 24 drugs with rigorous validation and testing using three datasets. The sequence features were represented as a large number of 10-mers and a good performance of an average MAE less than 1 was obtained from the models. The performance comparison of top-ranked 10-mers and all 10-mers highlighted the importance of feature selection. The top-ranked 10-mers can be further mapped to the genome for identifying genes relevant to antimicrobial susceptibility.

The proposed random forest algorithm is a non-linear learning method whose prediction is derived from complex 10-mer-based rules of an ensemble of decision trees. Since AMR can involve multiple genes simultaneously, the proposed method can provide better performance than traditional AMR gene-based methods. However, the identified 10-mers may not be directly linked to a specific phenotype making the interpretation of the associations difficult. The tradeoff between interpretability and prediction performance is a well-known issue. Machine learning methods can provide good performance [28], while the traditional AMR gene identification tool provides good interpretability. Other tools such as DBG-WAS [29], while not designed to build predictors for maximizing the prediction performance of phenotypes,



may also be utilized to study the associations of k-mers and phenotypes. Furthermore, while out of the scope of this study, as the phylogenetic information can benefit the genome-wide association study [30], the information may be further engineered as new features and evaluated for its contribution to MIC prediction.

As the utilized dataset was obtained from a regular surveillance program on healthy poultry and swine, the data imbalanced issues were expected and that may not hamper the utilization of the model due to the nature of the model for predicting MIC values rather than susceptibility. However, prediction performance of susceptibility classification should be interpreted with care. For example, enrofloxacin was associated with only two, one and zero resistant isolates in the training, validation and test dataset, respectively. The predictor learned more from the MIC distribution of the majority class of susceptible isolates and predicted the susceptible isolates well. As a result, the accuracy for enrofloxacin is notably high, and this can primarily be attributed to the majority class.

For surveillance purposes, the genome sequence-based machine learning methods could be utilized to monitor the difference between predicted and experimental MIC, where a large difference might be worthy of investigation on the emerging genomic determinants. This study presented a successful machine learning-based MIC prediction method utilizing genomic and phenotypic data obtained from surveillance programs in Taiwan. The incorporation of future data is expected to further improve the prediction performance of the models.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13567-023-01141-5>.

**Additional file 1.** The OOB validation performance using the training dataset.

**Additional file 2.** The validation performance using the validation dataset.

**Additional file 3.** The test performance using the independent test dataset.

**Additional file 4.** The OOB validation performance using the integrated dataset.

**Additional file 5.** The top-ranked k-mers for each drug.

**Additional file 6.** The OOB validation performance using the training dataset and top k-mers.

**Additional file 7.** The validation performance using the validation dataset and top k-mers.

**Additional file 8.** The test performance using the independent test dataset and top k-mers.

**Additional file 9.** The OOB performance using the integrated dataset and top k-mers.

**Additional file 10.** Prediction performance of an XGBoost-based model [16] using the independent test dataset.

**Additional file 11.** Prediction performance of ResFinder using the independent test dataset.

## Acknowledgements

We thank the data provider of the Bureau of Animal and Plant Health Inspection and Quarantine of Taiwan.

## Authors' contributions

CCW, YTH, TRJ and CWT conceived the idea. CYC, PYC and CWT implemented the program. CCW, YTH, CYC, ZWC, SLH, PYC, TRJ and CWT analyzed the data. All authors read and approved the final manuscript.

## Funding

This work was supported by Ministry of Science and Technology of Taiwan (MOST-107-2221-E-400-004-MY3, MOST-110-2221-E-400-004-MY3) and Bureau of Animal and Plant Health Inspection and Quarantine of Taiwan (109AS-8.8.1-BQ-B1, 110AS-5.6.1-BQ-B1, 111AS-5.6.1-BQ-B1, and 112AS-5.1.3-BQ-B1).

## Availability of data and materials

The data that support the findings of this study are available from the Bureau of Animal and Plant Health Inspection and Quarantine of Taiwan but restrictions apply to the availability of these data, which were used under the contract numbers 109AS-8.8.1-BQ-B1, 110AS-5.6.1-BQ-B1 and 111AS-5.6.1-BQ-B1 for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Bureau of Animal and Plant Health Inspection and Quarantine of Taiwan.

## Declarations

### Competing interests

The views presented in this article do not necessarily reflect current or future opinions or policies of the Bureau of Animal and Plant Health Inspection and Quarantine of Taiwan.

Received: 10 October 2022 Accepted: 13 January 2023

Published online: 06 February 2023

## References

- Li M, Havelaar AH, Hoffmann S, Hald T, Kirk MD, Torgerson PR, Devleeschauwer B (2019) Global disease burden of pathogens in animal source foods, 2010. *PLoS One* 14:e0216545
- Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM (2010) The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clin Infect Dis* 50:882–889
- Kurtz JR, Goggins JA, McLachlan JB (2017) Salmonella infection: interplay between the bacteria and host immune system. *Immunol Lett* 90:42–50
- Jajere SM (2019) A review of *Salmonella enterica* with particular focus on the pathogenicity and virulence factors, host specificity and antimicrobial resistance including multidrug resistance. *Vet World* 12:504–521
- World Health Organization (2018) WHO Advisory group on integrated surveillance of antimicrobial resistance (AGISAR): report of the 7<sup>th</sup> meeting, October 2016. Raleigh, USA
- Antimicrobial Resistance Collaborators (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* 399:629–655
- Sharma C, Rokana N, Chandra M, Singh BP, Gulhane RD, Gill JPS, Ray P, Puniya AK, Panwar H (2018) Antimicrobial resistance: its surveillance, impact, and alternative management strategies in dairy animals. *Front Vet Sci* 4:237
- Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF (2019) Using genomics to track global antimicrobial resistance. *Front Public Health* 7:242
- Köser CU, Ellington MJ, Peacock SJ (2014) Whole-genome sequencing to control antimicrobial resistance. *Trends Genet* 30:401–407

10. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, Grundman H, Hasman H, Holden MTG, Hopkins KL, Iredell J, Kahlmeter G, Köser CU, MacGowan A, Mevius D, Mulvey M, Naas T, Peto T, Rolain JM, Samuelsen Ø, Woodford N (2017) The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin Microbiol Infect* 23:2–22
11. Besser JM (2018) *Salmonella* epidemiology: a whirlwind of change. *Food Microbiol* 71:55–59
12. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen AV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran HK, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG et al (2020) CARD 2020: antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 48:D517–525
13. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK, Stingl K, Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum MF, Duggett NA, Kempf I, Nykäsenoja S, Olkkola S, Wiecek K, Amaro A, Clemente L, et al. (2020) ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* 75:3491–3500
14. Lau HJ, Lim CH, Foo SC, Tan HS (2021) The role of artificial intelligence in the battle against antimicrobial-resistant bacteria. *Curr Genet* 67:421–429
15. Anahar MN, Yang JH, Kanjilal S (2021) Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *J Clin Microbiol* 59:e0126020
16. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, Tyson GH, Zhao S, Davis JJ (2019) Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J Clin Microbiol* 57:e01260–e1318
17. Steinkey R, Moat J, Gannon V, Zovoilis A, Laing C (2020) Application of artificial intelligence to the in silico assessment of antimicrobial resistance and risks to human and animal health presented by priority enteric bacterial pathogens. *Can Commun Dis Rep* 46:180–185
18. Mouton JW, Meletiadi J, Voss A, Turnidge J (2018) Variation of MIC measurements: the contribution of strain and laboratory variability to measurement precision. *J Antimicrob Chemother* 73:2374–2379
19. Hicks AL, Wheeler N, Sánchez-Busó L, Rakeman JL, Harris SR, Grad YH (2019) Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLoS Comput Biol* 15:e1007349
20. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, Maojo V, Pazos A, Fernández-Lozano C (2021) A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J* 19:4538–4558
21. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma Oxf Engl* 30:2114–2120
22. Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 13:e1005595
23. Hendriksen RS, Kjeldgaard JS (2021) Protocol for whole genome sequencing and bioinformatic analysis of bacterial isolates related to the EU monitoring of antimicrobial resistance. [https://www.eurl-ar.eu/CustomerData/Files/Folders/34-wgs/628\\_protocol-for-wgs-v2-2.pdf](https://www.eurl-ar.eu/CustomerData/Files/Folders/34-wgs/628_protocol-for-wgs-v2-2.pdf). Accessed 13 Jan 2023.
24. Gut AM, Vasiljevic T, Yeager T, Donkor ON (2018) *Salmonella* infection—prevention and treatment by antibiotics and probiotic yeasts: a review. *Microbiol Read Engl* 164:1327–1344
25. Kokot M, Dlugosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. *Bioinforma Oxf Engl* 33:2759–2761
26. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L (2018) Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS Comput Biol* 14:e1006258
27. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, Chlenski P, Conrad N, Dickerman A, Dietrich EM, Gabbard JL, Gerdes S, Guard A, Kenyon RW, Machi D, Mao C, Murphy-Olson D, Nguyen M, Nordberg EK, Olsen GJ, Olson RD, Overbeek JC, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomas C, VanOeffelen M, Vonstein V, Warren AS, et al. (2020) The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res* 48:D606–612
28. Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP (2020) Reaching the end-game for GWAS: Machine learning approaches for the prioritization of complex disease loci. *Front Genet* 11:350
29. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, Jacob L (2018) A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. *PLoS Genet* 14:e1007758
30. Coll F, Gouliouris T, Bruchmann S, Phelan J, Raven KE, Clark TG, Parkhill J, Peacock SJ (2022) PowerBacGWAS: a computational pipeline to perform power calculations for bacterial genome-wide association studies. *Commun Biol* 5:266

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

